

# Ancestral sequence reconstruction with Maximum Parsimony

Lina Herbst · Mareike Fischer

Received: date / Accepted: date

**Abstract** One of the main aims in phylogenetics is the estimation of ancestral sequences based on present-day data like, for instance, DNA alignments. One way to estimate the data of the last common ancestor of a given set of species is to first reconstruct a phylogenetic tree with some tree inference method and then to use some method of ancestral state inference based on that tree. One of the best-known methods both for tree inference as well as for ancestral sequence inference is Maximum Parsimony (MP). In this manuscript, we focus on this method and on ancestral state inference for fully bifurcating trees. In particular, we investigate a conjecture published by Charleston and Steel in 1995 concerning the number of species which need to have a particular state, say  $a$ , at a particular site in order for MP to unambiguously return  $a$  as an estimate for the state of the last common ancestor. We prove the conjecture for all even numbers of character states, which is the most relevant case in biology. We also show that the conjecture does not hold in general for odd numbers of character states, but also present some positive results for this case.

**Keywords** Maximum parsimony · Fitch algorithm · ancestral sequence reconstruction

## 1 Introduction

Reconstructing data, like e.g. DNA sequences, of common ancestors of species living today is one of the main challenges of modern phylogenetics. Ancestral sequence reconstruction is a growing research field as information about the genetic makeup of ancestors is important in various areas, for instance in drug design, protein function investigation and comparative genomics [7, 12, 13]. Moreover, ancestral sequence estimation is crucial for understanding the evolution of today's species.

One possible way to estimate ancestral sequences is to consider the underlying evolutionary tree of the species of interest and to use Maximum Parsimony (MP) to infer ancestral sequences based on this tree. Some reasons that make MP interesting is that it does not require estimates of branch lengths on the input tree and that as a purely combinatorial method, it is a very

---

L. Herbst

Institute for Mathematics and Computer Science, Greifswald University, Walther-Rathenau-Str. 47, 17489 Greifswald, Germany

E-mail: herbstl@uni-greifswald.de

M. Fischer

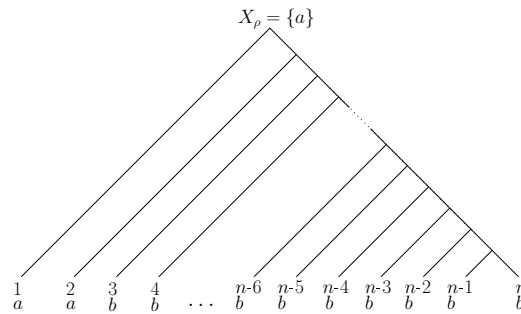
Institute for Mathematics and Computer Science, Greifswald University, Walther-Rathenau-Str. 47, 17489 Greifswald, Germany

E-mail: email@mareikefischer.de

fast method of ancestral sequence reconstruction [2, 5]. In this manuscript, we are interested in the last common ancestor of all species under investigation, not the ancestors of any subgroups. This means that we can use the famous Fitch algorithm [2] to infer *all* possible ancestral states (we will explain this more in-depth in the following section). For now just recall that the Fitch algorithm goes from the leaves to the root in a binary tree, i.e. in a tree where all nodes have exactly two children, and assigns each parental node a state set based on the state sets of its children. If the children's state sets intersect, this intersection is taken as the estimated state set for the ancestor. Otherwise, the union of the two sets is taken. For instance, in Figure 1, coming from the lowermost leaf, most inner nodes will be assigned  $\{b\}$ , but the parent of leaf 2 will be assigned  $\{a, b\}$ , because set  $\{a\}$  coming from leaf 2 and set  $\{b\}$  coming from leaves  $3, \dots, n$  (where  $n$  is the number of taxa) do not intersect. Then, root  $\rho$  is assigned  $\{a\}$  as the intersection of set  $\{a\}$  coming from leaf 1 and set  $\{a, b\}$  coming from leaves  $2, \dots, n$ , intersect. This means that the root of the tree will unambiguously be assigned  $a$  in this example.

However, MP cannot always make unambiguous decisions about the root sequence. Sometimes, for a particular site, MP might be indecisive between, say, the two DNA nucleotides  $A$  and  $C$  (note that our study is not restricted to the DNA alphabet, but can rather be applied to all kinds of alphabets, which is why we use lowercase letters in the following). Of course, for ancestral sequence reconstruction the question when the assignment is unique, i.e. when the decision is unambiguous, is of the utmost importance. In this regard, the following question arises: How many present-day species, i.e. how many leaves of the tree, need to be in a certain state, say  $a$ , such that the Maximum Parsimony estimate of the last common ancestor is also  $a$ ? Mathematically speaking, this question is equivalent to the question how many leaves need to be assigned state  $a$  (or 'colored' with 'color'  $a$ ), such that  $a$  is the unambiguous root state estimate in the underlying phylogenetic tree.

It can be easily seen that the answer to this question depends on the tree shape – for instance, the number of taxa; but also the height of the tree plays a crucial role. For instance, consider again the so-called caterpillar tree, which is depicted in Figure 1. This is an extreme case: As we have explained above, coloring two leaves, namely the uppermost two, with  $a$  is always sufficient to give an unambiguous root state estimate of  $a$ , even if all other leaves, which may be the vast majority, are in a different state. So for the caterpillar, two leaves in state  $a$  are always sufficient to give an unambiguous  $a$  root state estimate – regardless of the number of taxa under consideration.



**Fig. 1** A caterpillar tree, whose last common root is assigned  $\{a\}$  by just assigning two leaves  $a$ . Note that this construction works even if the total number of taxa in the tree is large.

On the other hand, if the number of species is  $n = 2^k$  for some natural number  $k$ , then you can also consider the so-called fully resolved and balanced tree, which is also often referred to as the fully bifurcating tree. The case  $k = 3$ , i.e.  $n = 8$ , is depicted in Figure 2. This tree turns out to be the opposite extreme: As all leaves are separated from the root by the same number of edges, it is clear that no constant number of leaves can have such a strong impact as in the caterpillar

case, but that instead the number of leaves which need to be assigned  $a$  in order for MP to return  $a$  as the unambiguous root state estimate will grow as  $k$  (and thus  $n$ ) grows. So in this regard, the fully bifurcating tree is of the utmost importance – once you know the minimal number of leaves that need to be assigned  $a$  in this particular tree in order for Maximum Parsimony to unambiguously return  $a$  as the root state estimate, you can use this number as an upper bound for *any* given tree on  $n = 2^k$  taxa.

So the present manuscript is concerned with the problem of finding the minimal number of leaves that need to be assigned  $a$  such that the MP root state estimate is  $\{a\}$  when the underlying tree is fully bifurcating. This problem has first been addressed in [1], where the authors stated a recursive formula which used the Fibonacci numbers and which explicitly calculates the desired number in case the alphabet under consideration consists only of  $r = 2$  states. So biologically, this result would only be applicable to binary data like for instance transversions versus transitions, but not to data like RNA or DNA (four states), proteins (20 states) or codons (64 states). However, in their manuscript the authors stated a conjecture for general alphabets of arbitrary size  $r$ . This conjecture inspired our work. In fact, in the present manuscript we first prove the conjecture for the case where  $r$  is even (which makes the result applicable to biological data like DNA, RNA, proteins or codons as stated above). We then analyze the more involved case where  $r$  is odd and show why the conjecture does not generally hold in this case, but also show some positive results for this case, like a simplification to only two different cases that need to be considered when  $r = 3$ .

## 1.1 Preliminaries

### 1.1.1 Basic definitions and notation

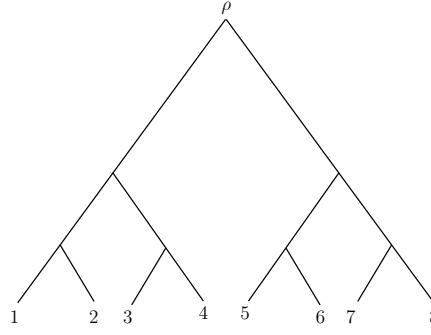
Before we can present our results, we first have to introduce some basic concepts. We start with some definitions. Recall that a *rooted binary phylogenetic tree on a taxon set  $L$*  (also sometimes referred to as a *rooted binary phylogenetic  $L$ -tree*) is a connected, acyclic graph in which the vertices of degree at most 1 are called *leaves*, and in which there is exactly one node  $\rho$  of degree 2, which is referred to as *root*, and all other non-leaf nodes have degree 3. Moreover, in a rooted binary phylogenetic  $L$ -tree the leaves are bijectively labelled by the elements of  $L$ . Note that in the special case where there is only one node, this node can be regarded as a leaf *and* as a root at the same time, i.e. this is the only case where the root does not have degree 2.

Throughout this paper, when we refer to trees, we always mean rooted binary phylogenetic  $L$ -trees, and we assume without loss of generality that  $L = \{1, \dots, n\}$ . Note all nodes on the path from a node  $v$  to the root  $\rho$  (except for  $v$ ) are called *ancestors* of  $v$ , and  $v$  is a *descendant* of all these nodes. Whenever the path from  $v$  to an ancestor  $u$  consists of only one edge,  $u$  is called *direct ancestor* of  $v$ , and  $v$  is called *direct descendant* of  $u$ .

Now let  $k \in \mathbb{N}$ . Then a *fully bifurcating phylogenetic tree of height  $k$* ,  $T_k$ , is a rooted binary phylogenetic tree which has exactly  $n = 2^k$  leaves and height  $k$ . Here, *height* refers to the number of edges between the root and each leaf (note that this number is uniquely determined for rooted binary trees with  $n = 2^k$  leaves). Figure 2 depicts the case  $n = 2^3 = 8$ , i.e. tree  $T_3$ . Note that for  $k = 0$ , i.e.  $n = 2^0 = 1$ , the corresponding fully bifurcating phylogenetic tree  $T_0$  consists of only one node, which is at the same time its root and only leaf.

Now that we introduced tree  $T_k$ , which we will investigate in this manuscript, we need to introduce two more concepts: the kind of data we will map onto the leaves of the tree as well as the method with which we will then estimate ancestral data like in particular the data of the root of the tree.

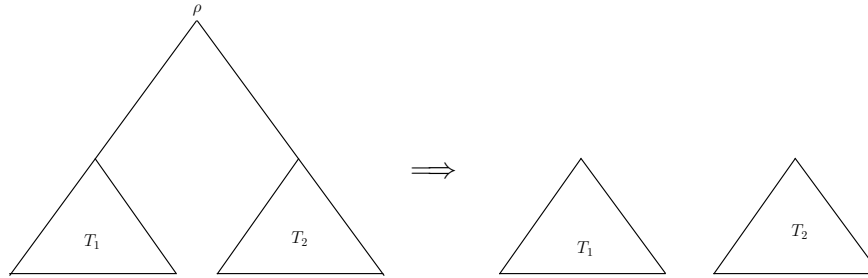
We first consider the data. Assume we have an alphabet  $R$  with  $|R| = r$  so-called *character states*, like e.g. the DNA alphabet  $R = \{A, C, G, T\}$ , where  $r = 4$ . Then a *character* on a taxon set  $L$  is just a function  $f$  from  $L$  to  $R$ , i.e.  $f : L \rightarrow R$ . If  $|f(L)| = \tilde{r}$ , then  $f$  is called  $\tilde{r}$ -state



**Fig. 2**  $T_3$  is a fully bifurcating phylogenetic tree of height 3 with  $2^3 = 8$  leaves.

character. Note that as we consider  $L = \{1, 2, \dots, n\}$ , we often write  $f = f(1)f(2)\dots f(n)$  instead of listing  $f(1)$ ,  $f(2)$ , etc. explicitly. For instance, Figure 1 shows an example of character  $f = aab\dots b$  mapped onto the leaves of a rooted binary phylogenetic tree. Note that given real data like e.g. a DNA alignment, a character can just be regarded as a column in the alignment, because each species is assigned one character state (however, note that in our case, alignments are considered to be gap-free). In this context, biologists also often refer to characters as *sites*.

We now want to turn our attention to the method which we will use to estimate ancestral sequence data, i.e. sequences of ancestral states. Given a rooted binary phylogenetic tree  $T$ , the *standard decomposition* of  $T$  is the decomposition of  $T$  into its two maximal rooted pending subtrees [4]. Informally speaking, as the root  $\rho$  has degree 2, this means that the standard decomposition considers the two subtrees directly below  $\rho$ . An illustration of this decomposition is given by Figure 3. Note that the fully bifurcating tree  $T_k$ , with which this manuscript is concerned, can be decomposed into two trees of shape  $T_{k-1}$ , which in turn can be decomposed into two  $T_{k-2}$  trees and so forth.



**Fig. 3** The standard decomposition of a rooted binary phylogenetic tree. The tree is decomposed into its two maximal rooted subtrees  $T_1$  and  $T_2$ .

A frequently used method for ancestral state reconstruction is *Maximum Parsimony*, or *MP* for short [5, Section 5]. Given a tree and a character, MP seeks to find possible state assignments to all internal nodes such that the number of edges in the so-called *change set* is minimized. The change set consists of all edges of  $T$  whose two endpoints are assigned different states. The minimal number of edges in the change set for a given tree  $T$  and a given character  $f$  is called *Maximum Parsimony score* (or *MP score* for short) of  $f$  on  $T$ . The MP score of an entire alignment, i.e. of a sequence of characters on  $T$ , is then simply the sum of all scores induced by the individual characters. This is why we investigate only individual characters in this manuscript rather than entire alignments.

Calculating the MP score for a given tree and a given character can be done in polynomial time using the famous *Fitch algorithm* [2], which in turn is based on *Fitch's parsimony operation*, which we define next.

**Definition 1** (*Fitch's parsimony operation*)

Let  $R$  be a nonempty finite set and let  $A, B \subseteq R$ . Then, *Fitch's parsimony operation*  $*$  is defined by

$$A * B := \begin{cases} A \cap B, & \text{if } A \cap B \neq \emptyset, \\ A \cup B, & \text{otherwise.} \end{cases}$$

This means that whenever the intersection of sets  $A$  and  $B$  is not empty,  $A * B$  is set to be equal to this intersection. Otherwise, the union is taken instead of the intersection.

Using this operation, the *Fitch algorithm* [2] works as follows (note that what we call the Fitch algorithm is in fact only one phase of the algorithm described in [2], but this is the only part required to estimate possible root states and is thus sufficient for the present manuscript). Assume we have a rooted binary tree  $T$  with leaf set  $L$  and a character  $f : L \rightarrow R$ . Now the vertices of  $T$  are recursively assigned non-empty subsets of  $R$ . First, each leaf is assigned the set consisting of the state assigned to it by  $f$ . Then all other vertices  $v$ , whose two direct descendants have already been assigned a subset, say  $A$  and  $B$ , are assigned the set  $A * B$ . This step is continued upwards along the tree until the root  $\rho$  is assigned a subset, which is denoted by  $X_\rho$ . The sets assigned to each node represent possible state sets which lead to assignments of states to all internal nodes, so-called *extensions of character  $f$* , such that the number of edges in the change set is minimized and thus the parsimony score is realized on  $T$  given  $f$ . Note that Fitch described two more phases [2]. One of them only derives extensions once the state sets for each node are given. However, the other one adds more states to some internal node sets as the procedure described above might not recover all possible states for all internal nodes [6]. However, note that the states suggested for the root  $\rho$  by the described algorithm do not require any correction [2]. Therefore, as the present manuscript is mainly concerned with  $\rho$ , the described procedure is sufficient and no corrections need to be made to  $X_\rho$ .

We are now in the position to turn our attention to the main question of our manuscript: Given a fully bifurcating phylogenetic tree  $T_k$  on taxon set  $L$  and height  $k$ , as well as an alphabet  $R$  with  $|R| = r$ , some set  $A \subseteq R$  and  $a \in A$ , what is the minimal number of leaves which must be assigned  $a$  such that  $X_\rho = A$ ? Given  $r \geq 2$  and  $|A| \leq 2^k$ , we denote this minimum by  $f_{k,r}^A$ . For simplicity, let  $f_{k,r} := f_{k,r}^{\{a\}}$ .

With the standard decomposition of rooted binary trees and Fitch's algorithm, it can be easily seen that

$$f_{k+1,r}^A = \min_{\substack{B, C \subseteq R, \\ |B|, |C| \leq 2^k}} \{f_{k,r}^B + f_{k,r}^C : B * C = A\}, \quad (1)$$

where  $*$  is the parsimony operation.

Note that the minimum is taken over all possible subsets  $B$  and  $C$  of  $R$ , which could be assigned to the direct descendants of  $\rho$  and which would then be used by Fitch's operation to calculate  $X_\rho$ . Whenever there is no ambiguity, we omit the subscripts and only refer to  $\min\{f_{k,r}^B + f_{k,r}^C\}$ . The main aim of this manuscript is to find a way to calculate this minimum without having to consider all possible sets  $B$  and  $C$ . Our work was motivated by the fact that in [1] it was shown by M. Steel and M. Charleston that  $f_{k,2}$  equals the  $(k+1)$ th Fibonacci number; so  $f_{k,2}$  can be calculated without considering all possible state assignments to the nodes of  $T_{k-1}$ . However, the authors did not consider the general case  $r > 2$ , but instead stated the following conjecture for this case which we seek to answer in the present manuscript.

**Conjecture 1 (Charleston and Steel [1])**

For  $T_k$  and  $r \geq 2$  we have:

$$f_{k,r} = \begin{cases} f_{k-p,r} + f_{k-p-1,r} & \text{when } r = 2p, p \in \mathbb{N}_{\geq 1}, \\ 2 \cdot f_{k-p,r} & \text{when } r = 2p - 1, p \in \mathbb{N}_{\geq 2}. \end{cases} \quad (2)$$

Note that Conjecture 1 is only defined for  $k \geq p + 1$  if  $r = 2p$  and  $k \geq p$  if  $r = 2p - 1$ , because  $f_{k,r}$  is only defined for  $k \geq 0$  as  $k$  is the height of the underlying tree  $T_k$ . Also note that  $f_{k,r}^A$  is only defined whenever  $|A| \leq 2^k$ , because by the Fitch operation you can never get more states into the root estimate set than there are leaves (as each leaf might be assigned a different state, which would lead to a root state set of size  $2^k$ , but more than that is not possible). So in the following, even if we do not state explicitly that our theorems are only true for the regions where  $f_{k,r}^A$  is defined, we will implicitly always assume that only these regions are considered. In particular, our proofs only consider these cases.

It is important to note that the initial conditions required for the start of the recursion are not specified in Conjecture 1. In the following we show that Conjecture 1 is not valid in general, because the case where  $r$  is odd is problematic unless the initial conditions are chosen in a specific way, and we also show relevant examples to illustrate this result.

Before we present our results, we want to introduce one last piece of notation, which simplifies all following equations a lot: In the following,  $A_i$  always denotes a subset of our alphabet  $R$  that contains  $i$  elements in total, one of which is  $a$ , i.e.  $A_i \subseteq R$  with  $a \in A_i$  and  $|A_i| = i$ . So in particular, set  $A_i$  contains  $a$  for all  $i$ . On the other hand, let  $D_j \subseteq R \setminus A_i$  with  $|D_j| = j$  and  $j = 1, \dots, r - i$ . In particular,  $D_j$  does not contain  $a$  for any  $j$ , and  $i + j = r = |R|$ .

## 2 Results

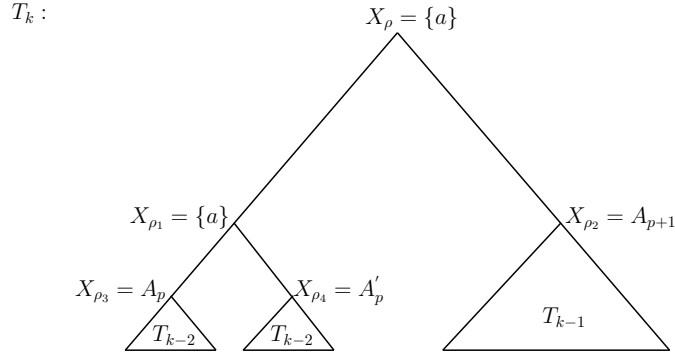
### 2.1 First insights into the case where $r$ is odd

Before we present our main result, which is concerned with the case in which the size  $r$  of the alphabet under consideration is even, we first consider the case where  $r$  is odd and present an example for which Conjecture 1 fails when the initial conditions are not chosen in a particular way.

**Observation 1** *Let  $p \in \mathbb{N}_{\geq 2}$  and  $r = 2p - 1$ . Then for  $k = p + 1$ , Conjecture 1 suggests  $f_{p+1,r} = 2 \cdot f_{(p+1)-p,r} = 2 \cdot f_{1,r} = 2 \cdot 2 = 4$ . This is due to the fact that in order for the root  $\rho$  of  $T_1$ , i.e. the tree that only consists of  $\rho$  and two leaves, to be assigned  $a$ , both leaves need to be in state  $a$ , so  $f_{1,r} = 2$  (this scenario is depicted on the left-hand side of Figure 7). However,  $f_{p+1,r} = 4$  is incorrect. In fact,  $f_{p+1,r} = 3$ .*

We will now prove Observation 1 and illustrate the construction presented in the proof subsequently with the explicit example where  $r = 3$ .

*Proof* Our proof strategy is depicted in Figure 4 and works as follows:



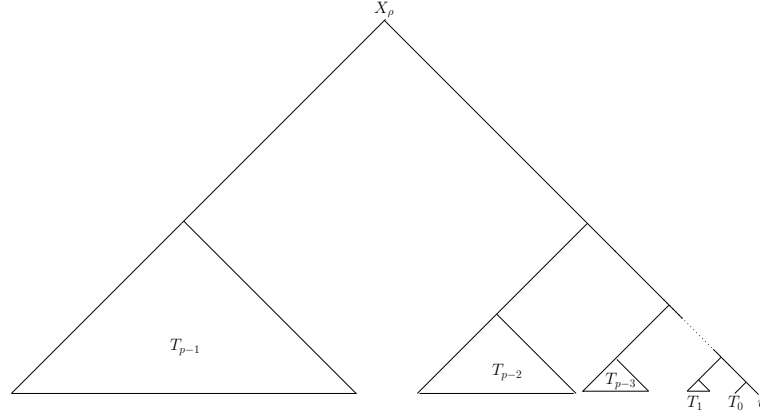
**Fig. 4** This figure illustrates the idea of the proof: We consider  $T_k$  and its standard decomposition into two trees  $T_{k-1}$ , for one of which we consider its standard decomposition of two trees  $T_{k-2}$ .

We consider tree  $T_k$  and its standard decomposition into two trees  $T_{k-1}$ . We will construct a character such that one of the two  $T_{k-1}$  trees is assigned  $A_{p+1}$  as MP root estimate, i.e.  $p+1$  distinct character states, one of which is  $a$ . For the other subtree  $T_{k-1}$ , we again consider its standard decomposition into two subtrees  $T_{k-2}$ , both of which we will assign sets  $A_p$  and  $A'_p$ , respectively, i.e. each subtree  $T_{k-2}$  gets assigned a set with  $p$  elements (one of which is  $a$ ) to its root. The trick is that as we have  $r = 2p - 1$  states, we can achieve the latter by letting  $A_p = \{a, c_1, \dots, c_{p-1}\}$  and  $A'_p = \{a, c_p, \dots, c_{2p-2}\}$ , where  $R = \{a, c_1, \dots, c_{2p-2}\}$  and thus  $|R| = r = 2p - 1$ , and so  $A_p \cap A'_p = \{a\}$ . Thus, the root of  $T_{k-1}$  would have set  $\{a\}$  as a root estimate, and  $T_k$  would have the root estimate  $\{a\} * A_{p+1} = \{a\} \cap A_{p+1} = \{a\}$  as required. We will present a construction that can achieve this by assigning  $a$  to only three leaves in total, one in one of the  $T_{k-1}$  subtrees and one in each  $T_{k-2}$  subtree of the other  $T_{k-1}$  subtree as described above.

So we first consider tree  $T_{k-1} = T_p$  and consider its subtrees as depicted in Figure 5. For this tree, we construct a character that only employs one  $a$  and leads to an MP root assignment of  $A_{p+1}$ , i.e. of a set with  $p+1$  elements, one of which is  $a$ . The construction is as follows: We assign leaf  $v$  with  $a$  and all leaves in subtrees  $T_0, \dots, T_{p-1}$  are all assigned only one of the  $r = 2p - 1$  character states, but each tree gets a unique one. So in total, the character uses  $p+1$  distinct character states, say  $a, c_1, \dots, c_p$ , and it can easily be seen that the Fitch algorithm will lead to the root assignment  $X_\rho = \{a, c_1, \dots, c_p\} = A_{p+1}$ .

Note that we can use the same construction to assign sets  $A_p$  and  $A'_p$  to trees  $T_{k-2}$ , where we can choose these two sets such that they intersect only in  $a$  as explained above.

So in total, we have used  $a$  three times and derive  $X_\rho = \{a\}$ . This already contradicts  $f_{p+1,r} = 4$  as suggested by Conjecture 1. Moreover, one can show that 3 is best possible, i.e. there is no character assigning  $a$  to only two leaves and giving  $X_\rho = \{a\}$  (calculation not shown). This completes the proof.  $\square$



**Fig. 5** The decomposition of the fully bifurcating tree  $T_p$  of height  $p$  as needed in the construction presented in the proof of Observation 1 and later in the proof of Lemma 2. In this construction, leaf  $v$  is assigned  $a$ ,  $T_0$  is assigned another state, and all other subtrees which are sketched here as triangles are assigned one state each such that all states used are pairwise distinct.

Note that the problem for  $r = 2p - 1 \geq 3$  as described in the construction of our proof of Observation 1 could be avoided if the recursion was started with  $f_{p,r} = 2$  and  $f_{p+1,r} = 3$  as initial conditions (rather than  $f_{1,r} = 2$  and  $f_{2,r} = 3$ ). So a careful choice of initial conditions is definitely important when  $r$  is odd, and for trees whose height is too small, the recursion is thus not applicable. However, note that the construction we used is not the only way to let Conjecture 1 fail when  $r$  is odd – in fact, there are also other examples, where the conjecture fails (result not shown). This is why in Section 2.3 we will analyze the case  $r = 3$  more in-depth and provide some simplifications concerning  $f_{k,3}$ .

We complete this section by providing an example in order to illustrate the construction presented in the proof of Observation 1.

*Example 1* Let  $R = \{a, b, c\}$ , i.e.  $r = 3$ . Let  $p = 2$ , and thus  $r = 2p - 1$  as required. By Conjecture 1 we would have for all  $k \geq 2$  that  $f_{k,3} = 2 \cdot f_{k-2,3}$ . Hence  $f_{3,3} = 2 \cdot f_{1,3} = 2 \cdot 2 = 4$ , since  $f_{1,r} = 2$  for all  $r$  (see Figure 7).

Consider  $T_3$  with the character  $f = abcabbca$  as depicted in Figure 6. This assignment of states to the leaves shows that  $f_{3,3} \leq 3$ . Note that an exhaustive search through all possible characters even confirms that  $f_{3,3} = 3$  (calculation not shown). In any case, this shows that Conjecture 1, which would lead to  $f_{3,3} = 4$ , is not correct.

So it is already obvious that calculating  $f_{k,r}$  recursively is not trivial and that Conjecture 1 does not in general hold when  $r$  is odd. However, the present manuscript has two main aims: We want to prove that the conjecture holds for the case where  $r$  is even, and we want to prove a recursion for the case where  $r = 3$ . Concerning the second aim, we are now already in the position to state the desired result, which we will prove subsequently in Section 2.3.

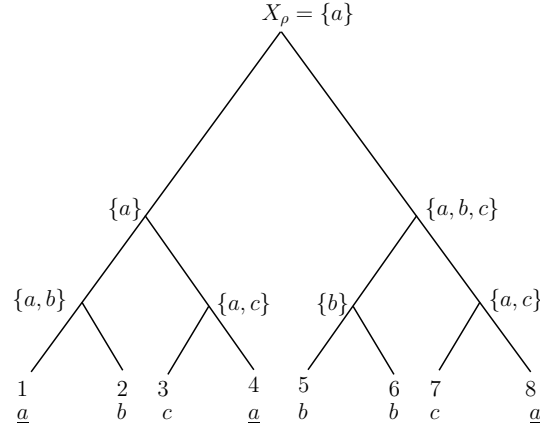
**Main Theorem 1** *Let  $T_k$  be a fully bifurcating tree of height  $k$  and let  $R$  be a set of character states with  $|R| = r = 3$ . Then for all  $k \geq 3$  we have*

$$f_{k+1,3} = 2 \cdot f_{k,3}^{A_2} = 2 \cdot f_{k-1,3}.$$

*Moreover  $f_{2,3} = 2$  and  $f_{3,3} = 3$ .*

Note that this theorem basically states that Conjecture 1 is correct for  $r = 3$  as long as the initial conditions are chosen in the way stated in the theorem. So even if we cannot make a general statement about the involved case where  $r$  is odd, this theorem characterizes the cases for which the conjecture holds when  $r = 3$ .





**Fig. 6**  $f_{3,3} \leq f_{2,3}^{\{a\}} + f_{2,3}^{\{a,b,c\}} = 2 + 1 = 3$ , since two leaves have to be assigned  $a$  to obtain  $\{a\}$  for the root of the left subtree of height 2 and one leaf have to be assigned  $a$  to obtain  $\{a, b, c\}$  for the root of the right subtree of height 2. Applying the parsimony operation for these two sets results in  $X_\rho = \{a\}$ .

The proof of this theorem requires some general properties of  $f_{r,k}^A$ , which will later also be needed to address the case where  $r$  is even. So now we first turn our attention to some general properties of  $f_{r,k}^A$ , some of which will be proved in the appendix.

## 2.2 General properties of $f_{k,r}^A$

We now investigate some general properties of  $f_{k,r}^A$ . These properties will be used later on to prove our main theorems. However, as we will point out, some properties of  $f_{k,r}^A$  are rather intuitive, even if they come with technical proofs. These proofs can be found in the appendix. Our main aim in this section is to simplify Equation (1) in the following sense: Rather than taking the minimum of all subsets  $B$  and  $C$  such that  $B, C \subseteq R$  and  $|B|, |C| \leq 2^k$ , we want to see if some sets can already be discarded as they can never lead to the minimum.

Some first properties of  $f_{k,r}^A$  are stated in the following lemma and can be derived from (1).

**Lemma 1** *Let  $A \subseteq R$ ,  $r = |R|$  and  $f_{k,r}^A$  be as defined in (1). Then, the following statements hold:*

- (i) *If  $a \in A \cap B$  and  $|A| = |B|$ , then  $f_{k,r}^A = f_{k,r}^B$ .*
- (ii) *If  $k = 0$ , then  $f_{0,r}^{\{a\}} = f_{0,r} = 1$  for all  $r \in \mathbb{N}$ .*
- (iii) *If  $k \geq 1$ , then  $f_{k,r}^{\{a\}} = f_{k,r} \geq 2$  for all  $r \in \mathbb{N}$ .*

A formal proof of Lemma 1 can be found in the appendix. However, note that the first two of the three statements of the lemma are not surprising: The first statement says that if you have two sets  $A$  and  $B$  which both contain  $a$  and have the same size, then you need to assign the letter  $a$  to equally many leaves in order to receive  $A$  or  $B$  as an MP root state estimate. This makes sense because the exact names of the states other than  $a$  in a set do not matter; what matters is rather the number of other states in the respective set. The second statement says that the fully bifurcating tree of height 0, i.e. a tree that consists only of one node, requires this one node to be assigned  $a$  if the root state estimate shall be  $\{a\}$  – this is clear as the root and the only leaf coincide in this case. The third statement is maybe the most interesting and least obvious one of the three statements. It says that if the root estimate shall be an unambiguous  $\{a\}$ , any fully bifurcating tree of height at least 1 requires at least two leaves that are assigned  $a$ .

Next we state a theorem which might seem obvious at first, too, but which is really crucial: If you have two sets  $A, B \subseteq R$ , which both contain  $a$ , and if you want to obtain the *smaller* set as an MP root state estimate, *more* leaves have to be assigned  $a$ .

**Theorem 1** *Let  $k \geq 1$  and  $A, B \subseteq R$  such that  $|A| \geq |B|$  with  $a \in A \cap B$  and  $|R| = r \geq 2$ . Then, we have:  $f_{k,r}^A \leq f_{k,r}^B$ .*

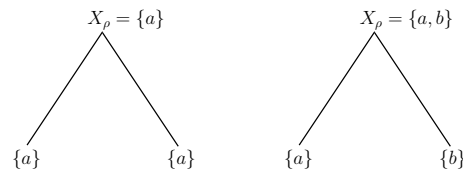
Note that, again, the statement of Theorem 1 is intuitively clear, because, roughly speaking, the smaller the set of the MP root state estimate is, the more Maximum Parsimony is convinced that  $a$  is a good root state candidate. For instance, if  $a$  is one state of only a few possible root states (i.e. many other states can be discarded as possible root state candidates), this outcome might be regarded as more relevant than if  $a$  is one state of many possible states (i.e. only a few other states can be discarded). In order to achieve this higher degree of certainty, more leaves with  $a$  are needed than if a smaller degree of certainty is sufficient.

We will now give a formal proof of Theorem 1, as not only the theorem itself, but also some aspects of the proof, are needed to establish subsequent results.

*Proof* If  $|A| = |B|$ , then  $f_{k,r}^A = f_{k,r}^B$  (see Lemma 1 (i)), so there remains nothing to show. Now assume  $|A| > |B|$ . Note that without loss of generality, we may assume  $B \subset A$ . This is due to the fact that if there were states in  $B$  that are not contained in  $A$ , i.e. if  $B \setminus (A \cap B) \neq \emptyset$ , then we can re-name these states using states of set  $A$  to get a set  $B'$  with  $|B'| = |B|$  and  $a \in B'$ . Then, by Lemma 1 (i),  $f_{k,r}^B = f_{k,r}^{B'}$ , but also  $B' \subset A$ . So we now assume that this re-naming of states in  $B$  has already taken place, so without loss of generality,  $B \subset A$ .

We proceed by induction on  $k$ .

For  $k = 1$  we have  $|A| \leq 2^k = 2$ , because the number of states that can occur in any MP root state estimate is clearly limited by the number of leaves (by Definition 1 of the parsimony operation), as no states that occur in none of the leaves can ever enter the root state estimate. Moreover, it can easily be seen that  $f_{1,r}^{\{a\}} = 2$  and  $f_{1,r}^{\{a,b\}} = 1$  hold for all  $r \geq 2$  (see Figure 7) and therefore Theorem 1 is true for  $k = 1$ .



**Fig. 7** Fully bifurcating trees of height 1 with  $X_\rho = \{a\}$  and  $X_\rho = \{a, b\}$  (w.l.o.g.  $A_2 = \{a, b\}$ ). To obtain  $X_\rho = \{a\}$  both leaves have to be assigned  $a$ . To obtain  $X_\rho = \{a, b\}$ , one leaf has to be assigned  $a$  and the other one  $b$ .

Now suppose Theorem 1 is true for  $k$ . Then we consider  $k+1$  and use Equation (1) to describe  $f_{k+1,r}^A$  as follows:

$$\begin{aligned}
 f_{k+1,r}^A &= \min \left\{ \begin{array}{ll} f_{k,r}^{A \cup S} + f_{k,r}^{A \cup \widehat{S}} : & S \cap \widehat{S} = \emptyset, S \cap A = \emptyset, \widehat{S} \cap A = \emptyset \\ & \text{and } |S| + |\widehat{S}| \leq r - |A|. \end{array} \right. \\
 &\quad \left( \text{where the minimum is taken over all } C_1, C_2, S, \widehat{S}, \text{ respectively} \right) \\
 &= \min \left\{ \begin{array}{ll} f_{k,r}^{A \cup S} + f_{k,r}^{A \cup \widehat{S}} : & S \cap \widehat{S} = \emptyset, S \cap A = \emptyset, \widehat{S} \cap A = \emptyset \\ & \text{and } |S| + |\widehat{S}| \leq r - |A|. \end{array} \right. \\
 &\quad \left( \text{since } a \notin C_2 \text{ and therefore } f_k^{C_2} = 0 \right) \\
 &= \min \left\{ \begin{array}{ll} f_{k,r}^{A \cup S} + f_{k,r}^{A \cup \widehat{S}} : & S \cap \widehat{S} = \emptyset, S \cap A = \emptyset, \widehat{S} \cap A = \emptyset \\ & \text{and } |S| + |\widehat{S}| \leq r - |A|. \end{array} \right. \\
 &\quad \left( \text{as Theorem 1 holds for } k \right). \tag{3}
 \end{aligned}$$

Analogously, we derive

$$\begin{aligned}
 f_{k+1,r}^B &= \min \left\{ \begin{array}{ll} f_{k,r}^{B \cup T} + f_{k,r}^{B \cup \widehat{T}} : & T \cap \widehat{T} = \emptyset, T \cap B = \emptyset, \widehat{T} \cap B = \emptyset \\ & \text{and } |T| + |\widehat{T}| = r - |B|. \end{array} \right. \\
 &\quad \left( \text{where the minimum is taken over } B_1, T, \widehat{T} \right) \\
 &= \min \left\{ \begin{array}{ll} f_{k,r}^{B_1} : & a \in B_1, B_1 \subset B \text{ and } |B_1| = |B| - 1. \end{array} \right.
 \end{aligned}$$

So for  $f_{k+1,r}^B$ , there are basically two cases, depending on whether this value is obtained by its two maximal pending subtrees in the standard decomposition by an intersection or a union. We consider these two cases for  $f_{k+1,r}^B$  separately.

**1<sup>st</sup> case:**  $f_{k+1,r}^B = f_{k,r}^{B_1}$  with  $a \in B_1$ ,  $B_1 \subset B$  and  $|B_1| = |B| - 1$ .

Furthermore, we have that  $A \subseteq R$  with  $B \subset A$ . Moreover, let  $C_1 \subset A$  with  $a \in C_1$  and  $|C_1| = |A| - 1$ . Note that then  $|B_1| < |C_1|$ . We conclude:

$$\begin{aligned}
 f_{k+1,r}^B &= f_{k,r}^{B_1} \\
 &\geq f_{k,r}^{C_1} && \text{as Theorem 1 holds for } k \\
 &\geq f_{k+1,r}^A && \text{by (3).}
 \end{aligned}$$

**2<sup>nd</sup> case:**  $f_{k+1,r}^B = f_{k,r}^{B \cup T} + f_{k,r}^{B \cup \widehat{T}}$  with  $T \cap \widehat{T} = \emptyset$ ,  $T \cap B = \emptyset$ ,  $\widehat{T} \cap B = \emptyset$  and  $|T| + |\widehat{T}| = r - |B|$ . Moreover, we have  $A \subseteq R$  and  $B \subset A$ . Now, as we have  $B \subset A$ , we can conclude  $B \cup T \subseteq A \cup T$  for all  $T$  and  $B \cup \widehat{T} \subseteq A \cup \widehat{T}$  for all  $\widehat{T}$ . Hence

$$\begin{aligned}
 f_{k+1,r}^B &= f_{k,r}^{B \cup T} + f_{k,r}^{B \cup \widehat{T}} \\
 &\geq f_{k,r}^{A \cup T} + f_{k,r}^{A \cup \widehat{T}} && \text{as Theorem 1 holds for } k \\
 &\geq f_{k+1,r}^A && \text{by (3).}
 \end{aligned}$$

So in both cases, we conclude  $f_{k+1,r}^A \leq f_{k+1,r}^B$ , which completes the proof.  $\square$

The proof of Theorem 1 (particularly Equation (3)) helps us directly to gain more insight into  $f_{k+1,r}^{A_i}$ , which we will summarize in the following corollary.

**Corollary 1** *The following recursions hold:*

$$\begin{aligned}
f_{k+1,r}^{A_r} &= f_{k+1,r}^R = \min\{f_{k,r}^{A_{r-1}}, 2 \cdot f_{k,r}^{A_r}\}, \\
f_{k+1,r}^{A_{r-1}} &= \min\{f_{k,r}^{A_{r-2}}, f_{k,r}^{A_{r-1}} + f_{k,r}^{A_{r-1} \cup D_1}\}, \\
f_{k+1,r}^{A_{r-2}} &= \min\{f_{k,r}^{A_{r-3}}, f_{k,r}^{A_{r-2}} + f_{k,r}^{A_{r-2} \cup D_2}, 2 \cdot f_{k,r}^{A_{r-2} \cup D_1}\}, \\
&\vdots \\
f_{k+1,r}^{A_1} &= f_{k+1} = \min\{f_{k,r}^{A_1} + f_{k,r}^{A_1 \cup D_{r-1}}, f_{k,r}^{A_1 \cup D_1} + f_{k,r}^{A_1 \cup D_{r-2}}, f_{k,r}^{A_1 \cup D_2} + f_{k,r}^{A_1 \cup D_{r-3}}, \dots, F^{A_1}\}, \\
\text{where } F^{A_1} &:= \begin{cases} f_{k,r}^{A_1 \cup D_{p-1}} + f_{k,r}^{A_1 \cup D_p} & \text{if } r = 2p \\ 2 \cdot f_{k,r}^{A_1 \cup D_{p-1}} & \text{if } r = 2p - 1 \end{cases}.
\end{aligned}$$

As Corollary 1 is a direct consequence of (3), we omit the formal proof. Instead, we turn our attention to another rather intuitive property of  $f_{k,r}^A$ : namely, that as  $k$  (and thus the tree under consideration) grows, the number of leaves needed for an MP root state estimate of  $A$  cannot decrease.

**Theorem 2** *For all regions where  $f_{k,r}^A$  and  $f_{k+1,r}^A$  are defined, the following inequality holds:  $f_{k,r}^A \leq f_{k+1,r}^A$ . That is,  $f_{k,r}^A$  is monotonically increasing in  $k$  for all  $A \subseteq R$ .*

The proof of Theorem 2 is presented in the appendix.

Considering (3), it is obvious that in order to calculate  $f_{k+1,r}^A$  recursively, one has to consider two cases, which is due to the fact that the MP root state estimate for  $T_{k+1}$  can come from a union or an intersection at the level of the two maximum pending subtrees of the standard decomposition. However, as the following crucial theorem shows, in the case where  $2 \leq |A| \leq r$ , the recursion can be drastically simplified.

**Theorem 3** *Let  $i = 2, \dots, r$ . Then for all  $k + 1 \geq i - 1$  for which  $f_{k+1,r}^{A_i}$ ,  $f_{k,r}^{A_{i-1}}$  and  $f_{k-(i-2),r}$  are defined, we have:*

$$f_{k+1,r}^{A_i} = f_{k,r}^{A_{i-1}} = f_{k-(i-2),r}.$$

Note that Theorem 3 is not only crucial in the sense that it allows us to omit a distinction between union and intersection – it even allows for a reduction of a root state estimate  $A$  with  $|A| \geq 2$  to the unambiguous root state estimate  $\{a\}$  but for a smaller tree. The proof of this theorem can be found in the appendix, but in brief the idea of the first equation stated in Theorem 3 can be summarized like this: root state estimate  $A_i$  can be achieved by considering the standard decomposition of  $T_{k+1}$  into the two subtrees  $T_k$ , one of which gets assigned  $A_{i-1}$  as a root state estimate, and the other one gets assigned a set, say  $C$ , with  $|C| = 1$  and with  $a \notin C$  and  $A_{i-1} \cap C = \emptyset$ . This way, the root estimate of  $T_{k+1}$  will be  $A_{i-1} \cup C = A_i$ . Note that in order to achieve root state estimate  $C$  for  $T_k$ , no leaf in  $T_k$  needs to be labelled  $a$  (e.g. all leaves in this tree can be labelled with the unique state in  $C$ , which is not  $a$ ). So  $f_{k,r}^C = 0$  and thus  $f_{k+1,r}^{A_i} = f_{k,r}^{A_{i-1}} + f_{k,r}^C = f_{k,r}^{A_{i-1}}$ . The second equation of Theorem 3 follows recursively.

Note that by Theorem 3 and Equation (3) we can also express  $f_{k+1,r}$  in terms of  $f_{k,r}$ . In particular, we have

$$f_{k+1,r} = \min\{f_{k,r}^{A_1} + f_{k,r}^{A_r}, f_{k,r}^{A_2} + f_{k,r}^{A_{r-1}}, f_{k,r}^{A_3} + f_{k,r}^{A_{r-2}}, \dots, F_1\} \quad (4)$$

$$\text{where } F_1 := \begin{cases} f_{k,r}^{A_p} + f_{k,r}^{A_{p+1}} & \text{if } r = 2p \\ 2 \cdot f_{k,r}^{A_p} & \text{if } r = 2p - 1 \end{cases}$$

$$= \min\{f_{k,r} + f_{k-r+1,r}, f_{k-1,r} + f_{k-r+2,r}, f_{k-2,r} + f_{k-r+3,r}, \dots, F_2\}, \quad (5)$$

$$\text{where } F_2 := \begin{cases} f_{k-p+1,r} + f_{k-p,r} & \text{if } r = 2p \\ 2 \cdot f_{k-p+1,r} & \text{if } r = 2p - 1 \end{cases}.$$

So in order for the MP root state estimate to be  $\{a\}$ , we still have to find the minimum over various cases. Note that the case  $f_{k+1,r} = F_2$  corresponds to Conjecture 1. This means that Conjecture 1 basically states that the other options can be discarded as they never give the minimum. As we have already seen in Observation 1, this is not generally correct, though.

Before we can finally prove Main Theorem 1, we need one more useful property of function  $f_{k,r}^A$ . In particular, we find that if we reduce the number of character states that are available, the number of leafs that need to be in state  $a$  in order for Maximum Parsimony to give a root state estimate of  $A_i$  (where  $a \in A_i$  and  $|A_i| = i$  as before) cannot decrease. Again, this is intuitively clear, because we have already seen in the construction presented in the proof of Observation 1 that we could reduce the number of leafs in state  $a$  by making use of as many alternative character states as possible.

**Theorem 4** *Let  $R$  be a finite set of character states with  $|R| = r \geq 2$  and  $a \in R$ , let  $2 \leq \tilde{r} \leq r$  and  $a \in A_i \subseteq R$  with  $i \in \{1, \dots, \tilde{r}\}$ . Then, for all regions where  $f_{k+1,r}^{A_i}$  and  $f_{k+1,\tilde{r}}^{A_i}$  are defined, we have:*

$$f_{k+1,r}^{A_i} \leq f_{k+1,\tilde{r}}^{A_i}.$$

A formal proof of Theorem 4 can again be found in the appendix. This concludes our summary of basic properties of  $f_{k,r}^A$ , which we will use subsequently to prove our main results. We first focus our attention again on the case where  $r$  is odd; in particular, on the special case where  $r = 3$ .

### 2.3 Results for 3-state characters

We now consider the case where  $|R| = r = 3$ . By Theorem 1 and Theorem 3, we already have for  $r = 3$  that

$$f_{k+1,3}^{A_3} = f_{k+1,3}^R = f_{k,3}^{A_2}, \quad (6)$$

$$f_{k+1,3}^{A_2} = f_{k,3}, \quad (7)$$

$$\begin{aligned} f_{k+1,3}^{A_1} &= f_{k+1,3} = \min\{2 \cdot f_{k,3}^{A_2}, f_{k,3} + f_{k,3}^R\} \\ &= \min\{2 \cdot f_{k-1,3}, f_{k,3} + f_{k-2,3}\}. \end{aligned} \quad (8)$$

We are now finally in the position to prove Main Theorem 1.

*Proof (Main Theorem 1)* First consider the statements for  $f_{2,3}$  and  $f_{3,3}$ . As the construction presented in Figure 6 shows,  $f_{3,3} \leq 3$ . Moreover, the left-hand side subtree in Figure 6 also shows that  $f_{2,3} \leq 2$ . An exhaustive search through all possible characters on  $2^2 = 4$  and  $2^3 = 8$  leafs, respectively, shows that in both cases, equality holds (calculations not shown).

So now we turn our attention to the case  $k \geq 3$ . Without loss of generality, we can assume that

$k$  odd and prove Theorem 1 separately for  $k$  and  $k + 1$ , where the latter then covers the case where the tree height is even.

We first show the following two statements by induction on  $k$ :

$$2 \cdot f_{k-1,3} < f_{k,3} + f_{k-2,3}, \quad (9)$$

$$2 \cdot f_{k,3} = f_{k+1,3} + f_{k-1,3}. \quad (10)$$

For  $k = 3$  we have  $f_{1,3} = 2$  (see Figure 7),  $f_{2,3} = 2$  and  $f_{3,3} = 3$  (initial conditions as explained above) and, by exhaustive search,  $f_{4,3} = 4$  (reached by character *abacbbbabaccccc* on  $2^4$  leafs; calculation not shown). This leads to

$$2 \cdot f_{2,3} = 4 < 5 = f_{3,3} + f_{1,3},$$

$$2 \cdot f_{3,3} = 6 = f_{4,3} + f_{2,3},$$

which is the base case of the induction. Now suppose (9) and (10) are true for all  $h$  with  $3 \leq h \leq k$  and  $h$  odd. Then with (8),  $f_{k+1,3}$  and  $f_{k+2,3}$  become

$$f_{k+1,3} = 2 \cdot f_{k-1,3}, \quad (11)$$

$$f_{k+2,3} = 2 \cdot f_{k,3} = f_{k+1,3} + f_{k-1,3}, \quad (12)$$

where the second equation in (12) is due to (10).

Now we show that (9) and (10) are also true for  $k + 2$ .

$$\begin{aligned} 2 \cdot f_{(k+2)-1,3} &= 2 \cdot f_{k+1,3} \\ &= 2 \cdot 2 \cdot f_{k-1,3} && \text{by (11)} \\ &< 2 \cdot (f_{k,3} + f_{k-2,3}) && \text{by (9)} \\ &= 2 \cdot f_{k,3} + 2 \cdot f_{k-2,3} \\ &= f_{k+2,3} + f_{k,3} && \text{by (12)}. \end{aligned}$$

Therefore (9) holds for all  $k \geq 3$  and  $k$  odd.

$$\begin{aligned} 2 \cdot f_{k+2,3} &= 2 \cdot (f_{k+1,3} + f_{k-1,3}) && \text{by (12)} \\ &= 2 \cdot f_{k+1,3} + 2 \cdot f_{k-1,3} \\ &= f_{k+3,3} + f_{k+1,3} && \text{by (9)}. \end{aligned}$$

Hence also (10) holds for all  $k \geq 3$  and  $k$  odd, so that in total, (11) and (12) are also true for all  $k \geq 3$  and  $k$  odd. Thus we have that  $f_{k+1,3} = 2 \cdot f_{k-1,3}$  holds for all  $k \geq 3$ . So together with Theorem 3, we have  $2 \cdot f_{k-1,3} = 2 \cdot f_{k,3}^{A_2}$ , which completes the proof.  $\square$

Recall that Main Theorem 1 basically gives a characterization of the cases in which Conjecture 1 holds in the case where  $r = 3$ ; namely by choosing appropriate initial conditions in order to avoid the scenario constructed in the proof of Observation 1. However, this result cannot easily be extended to the general case where  $r$  is odd. We will discuss this further in Section 3.

We now turn our attention to the case where  $r$  is even, which will turn out to be less involved as Conjecture 1 can be shown to hold in this case.

## 2.4 Results for all even numbers of character states

We are finally in a position to state our second main result.

**Main Theorem 2** *Let  $T_k$  be a fully bifurcating tree of height  $k$  and let  $R$  be a set of character states with  $|R| = r = 2p$  and  $p \in \mathbb{N}_{\geq 1}$ . Then for  $k \geq r$  and all regions where  $f_{k+1,r}$ ,  $f_{k,r}^{A_p}$ ,  $f_{k,r}^{A_{p+1}}$ ,  $f_{k-p+1,r}$  and  $f_{k-p,r}$  are defined, we have:*

$$f_{k+1,r} = f_{k-p+1,r} + f_{k-p,r}.$$

Note that Theorem 2 states that Conjecture 1 holds in the case where  $r$  is even and  $k \geq r$ , and this statement does not depend on a specific choice of initial conditions. Before we can present the proof of Main Theorem 2, we need an additional lemma, which helps to cover the case  $k < r$ , which is not considered in the theorem. Recall that we already have  $f_{0,r} = 1$  and  $f_{1,r} = 2$  for all  $r$  (by Lemma 1 (ii) and Figure 7). The following lemma investigates the case  $k < r$  further.

**Lemma 2** *For  $r = 2p$  with  $p \in \mathbb{N}_{\geq 1}$ , we have:*

1.  $f_{k,r} = 2$  for all  $k = 1, \dots, p$
2.  $f_{p+1,r} = 3$
3.  $f_{k,r} \leq 4$  for all  $p+1 < k \leq r$ .

The proof of this lemma can be found in the appendix. Note that the bound stated in Lemma 2 is actually tight, as it can be shown that while  $f_{5,6} = f_{6,8} = 3$ , we have  $f_{7,8} = 4$  (calculations not shown). Whether  $f_{k,r} = 3$  is possible for  $r = 2p \geq 10$  and  $p+1 < k \leq r$  has not yet been investigated.

We now state the proof of Main Theorem 2.

*Proof (Main Theorem 2)* Note that by the second equation of Theorem 3, we have  $f_{k-p+1,r} + f_{k-p,r} = f_{k,r}^{A_p} + f_{k,r}^{A_{p+1}}$ . We now show that for all  $r = 2p$  with  $p \in \mathbb{N}_{\geq 1}$  we have

$$f_{k,r}^{A_p} + f_{k,r}^{A_{p+1}} \leq f_{k,r}^{A_i} + f_{k,r}^{A_j} \quad (13)$$

for all  $1 \leq i \leq j \leq r$  and  $i+j = r+1$ . Together with (4) this will prove Main Theorem 2. The proof is by induction on  $k$ . For  $k = r$  we need to show that

$$f_{r,r}^{A_p} + f_{r,r}^{A_{p+1}} \leq \min\{f_{r,r}^{A_1} + f_{r,r}^{A_r}, f_{r,r}^{A_2} + f_{r,r}^{A_{r-1}}, f_{r,r}^{A_3} + f_{r,r}^{A_{r-2}}, \dots, f_{r,r}^{A_{p-1}} + f_{r,r}^{A_{p+2}}\}, \quad (14)$$

which is, by Theorem 3, equivalent to showing

$$f_{r-p+1,r} + f_{r-p,r} \leq \min\{f_{r,r} + f_{r-r+1,r}, f_{r-1,r} + f_{r-r+2,r}, \dots, f_{r-p+2,r} + f_{r-p-1,r}\}.$$

Since  $r = 2p$ , this inequality can be written as

$$f_{p+1,r} + f_{p,r} \leq \min\{f_{r,r} + f_{1,r}, f_{r-1,r} + f_{2,r}, \dots, f_{p+2,r} + f_{p-1,r}\}. \quad (15)$$

Moreover, by Lemma 2 we know that  $f_{1,r} = f_{2,r} = \dots = f_{p,r} = 2$  and thus (15) results in

$$f_{p+1,r} \leq \min\{f_{r,r}, f_{r-1,r}, \dots, f_{p+2,r}\}. \quad (16)$$

By Theorem 2 we have  $f_{p+1,r} \leq f_{p+2,r} \leq \dots \leq f_{r-2,r} \leq f_{r-1,r} \leq f_{r,r}$ . So altogether, (16) holds and thus also (14). This completes the base case of the induction.

Now assume that (13) holds for  $k$ , then with (4) and Theorem 3, we get

$$f_{k+1,r} = f_{k,r}^{A_p} + f_{k,r}^{A_{p+1}} = f_{k-p+1,r} + f_{k-p,r}. \quad (17)$$

Now we show that (13) also holds for  $k + 1$ .

$$\begin{aligned}
f_{k+1,r}^{A_p} + f_{k+1,r}^{A_{p+1}} &= f_{k-(p-2),r} + f_{k-(p+1-2),r} && \text{by Theorem 3} \\
&= f_{k-p+2,r} + f_{k-p+1,r} \\
&= f_{k-p+1,r}^{A_p} + f_{k-p+1,r}^{A_{p+1}} + f_{k-p,r}^{A_p} + f_{k-p,r}^{A_{p+1}} && \text{by (17)} \\
&\leq f_{k-p+1,r}^{A_i} + f_{k-p+1,r}^{A_j} + f_{k-p,r}^{A_i} + f_{k-p,r}^{A_j} && \text{by (13)} \\
&\quad \text{for all } i, j : 1 \leq i \leq j \leq r \text{ and } i + j = r + 1 \\
&= f_{k-p-(i-2),r} + f_{k-p-(j-2),r} + f_{k-p-(i-1),r} + f_{k-p-(j-1),r} && \text{by Theorem 3} \\
&= f_{k-(p-1)-i+1,r} + f_{k-(p-1)-j+1,r} + f_{k-(p-1)-i,r} + f_{k-(p-1)-j,r} \\
&= f_{k-i+1,r}^{A_p} + f_{k-j+1,r}^{A_p} + f_{k-i+1,r}^{A_{p+1}} + f_{k-j+1,r}^{A_{p+1}} && \text{by Theorem 3} \\
&= f_{k-i+1,r}^{A_p} + f_{k-i+1,r}^{A_{p+1}} + f_{k-j+1,r}^{A_p} + f_{k-j+1,r}^{A_{p+1}} \\
&= f_{k+1-i+1,r} + f_{k+1-j+1,r} && \text{by (17)} \\
&= f_{k-(i-2),r} + f_{k-(j-2),r} \\
&= f_{k+1,r}^{A_i} + f_{k+1,r}^{A_j} && \text{by Theorem 3.}
\end{aligned}$$

Therefore (13), holds for all  $r = 2p$  with  $p \in \mathbb{N}^+$ . Applying (13) to (4) and (5) completes the proof.  $\square$

### 3 Conclusion and Discussion

We considered the question how many leafs of a fully bifurcating tree must be labelled  $a$  in order for the Maximum Parsimony root state estimate to be  $a$ . In this regard, we could prove a conjecture by Steel and Charleston for the case where the number  $r$  of character states is even [1]. This is an interesting result because it not only proves the conjecture, but also generalizes the findings presented in [1], where the case  $r = 2$  was investigated. We proved that the case where  $r$  is even is simple in the sense that the conjecture holds no matter where the recursion is started. On the other hand, we also showed that when  $r$  is odd, counterexamples to the conjecture can be constructed under certain conditions. However, we were able to prove the conjecture for  $r = 3$  for a careful choice of initial conditions. The general case where  $r$  is odd is a topic of ongoing research.

The conjecture by Steel and Charleston and thus also our results are relevant as the fully bifurcating tree is a ‘worst case scenario’ for unambiguous Maximum Parsimony root state estimation: As all leafs are equally far away from the root (in the sense of the number of edges that separates them from the root), there is no leaf that has a higher impact on the root state estimate than others. We showed in Section 1 that the so-called caterpillar tree is the other extreme: Here, no matter how many leafs we label with  $b$ , two (particular) leafs labelled  $a$  will be sufficient to get an unambiguous MP root state estimate of  $\{a\}$ , whereas for the fully balanced tree, the number depends strongly on the size of the tree (in terms of the height and thus the number of leafs). Recall that in the literature, the fully bifurcating tree is often referred to as ‘balanced’ as opposed to the caterpillar tree, which is often thought of as ‘unbalanced’ [15]. Trees with  $2^k$  leafs (for some  $k \geq 1$  which are ‘in-between’ these two trees in terms of balance will also need at least 2 leafs to be in state  $a$ , like the caterpillar tree, and at most  $f_{k,r}$  leafs in state  $a$  like the fully bifurcating tree, in order for MP to give an unambiguous root state estimate of  $\{a\}$ . Thus, our results present an upper bound for all trees of size  $2^k$ . However, the exact numbers for other tree shapes as well as other tree sizes are still to be investigated.



**Acknowledgements** We thank Mike Steel for bringing this topic to our attention. The first author also thanks the Ernst-Moritz-Arndt-University Greifswald for the Landesgraduieretenförderung studentship, under which this work was conducted.

## 4 Appendix

*Proof (Proof of Lemma 1)* (i) Let  $A, B$  such that  $a \in A \cap B$  and  $|A| = |B|$ . Then  $A$  could be transformed into  $B$  by renaming all states not element of  $A \cap B$ . Then  $A = B$ , and this yields  $f_{k,r}^A = f_{k,r}^B$ .

(ii) Let  $k = 0$ . In this case our tree consists of one leaf, which is at the same time the root. This vertex have to be assigned  $a$  to obtain  $X_\rho = \{a\}$ . Hence  $f_{0,r} = 1$  for all  $r$ .

(iii) Let  $k \geq 1$ . Then

$$\begin{aligned} f_{k,r} &= \min\{f_{k-1,r}^B + f_{k-1,r}^C : B * C = \{a\}\} \\ &= \min\{f_{k-1,r}^B + f_{k-1,r}^C : B \cap C = \{a\}\}, \\ &\text{since } B \cup C \text{ would not result in } X_\rho = \{a\} \text{ as } B, C \neq \emptyset. \end{aligned}$$

Therefore,  $a \in B$  and  $a \in C$  which results in  $f_{k,r} \geq 2$ .  $\square$

*Proof (Theorem 2)* We have  $f_{0,r}^{\{a\}} = 1$  (by Lemma 1 (ii)) and  $f_{1,r}^{\{a\}} = 2$  (see Figure 7) for all  $r$ . Therefore, Theorem 2 is true for  $k = 0$ . We now prove Theorem 2 for  $k > 0$  by contradiction. So assume there is a  $\hat{k}$  such that for some set  $A$  we have  $f_{\hat{k}+1,r}^A < f_{\hat{k},r}^A$ . Choose  $k$  to be the smallest value of  $\hat{k}$  with this property, i.e. we have  $f_{k+1,r}^A < f_{k,r}^A$  for some set  $A$ , which we fix. Moreover, for all  $\tilde{k} < k$  and all sets  $\tilde{A}$  we have  $f_{\tilde{k},r}^{\tilde{A}} \geq f_{\tilde{k}-1,r}^{\tilde{A}}$ . Now let  $A_1, S, \hat{S}$  and  $f_{k+1,r}^A$  be as in the proof of Theorem 1. We now consider both cases for  $f_{k+1,r}^A$  for the fixed values  $k$  and  $A$  as chosen above.

**1<sup>st</sup> case:**  $f_{k+1,r}^A = f_{k,r}^{A_1}$  with  $a \in A_1$ ,  $A_1 \subset A$  and  $|A_1| = |A| - 1$ . Then  $f_{k,r}^{A_1} = f_{k+1,r}^A < f_{k,r}^A$  by the assumption on  $k$ . This leads to  $f_{k,r}^{A_1} < f_{k,r}^A$ , which contradicts Theorem 1, because  $|A_1| = |A| - 1 < |A|$ .

**2<sup>nd</sup> case:**  $f_{k+1,r}^A = f_{k,r}^{A \cup S} + f_{k,r}^{A \cup \hat{S}}$  with  $S \cap \hat{S} = \emptyset$ ,  $S \cap A = \emptyset$ ,  $\hat{S} \cap A = \emptyset$  and  $|S| + |\hat{S}| = r - |A|$ . Then

$$\begin{aligned} f_{k,r}^A &> f_{k+1,r}^A && \text{by assumption on } k \\ &= f_{k,r}^{A \cup S} + f_{k,r}^{A \cup \hat{S}} \\ &\geq f_{k-1,r}^{A \cup S} + f_{k-1,r}^{A \cup \hat{S}} && \text{by assumption on all } \tilde{k} < k, \text{ in particular } \tilde{k} = k - 1 \\ &\geq f_{k,r}^A && \text{by (3).} \end{aligned}$$

Thus  $f_{k,r}^A > f_{k,r}^A$ , which is a contradiction.

Therefore, both cases lead to contradictions, and thus, such a  $k$  cannot exist. This completes the proof.  $\square$

*Proof (Theorem 3)* We begin by proving

$$f_{k,r}^{A_{i-1}} \leq f_{k,r}^{A_i \cup S} + f_{k,r}^{A_i \cup \hat{S}} \quad (18)$$

for  $2 \leq i \leq r$  for all  $S, \hat{S} \subseteq R : S \cap \hat{S} = \emptyset$ ,  $S \cap A_i = \emptyset$ ,  $\hat{S} \cap A_i = \emptyset$  and  $|S| + |\hat{S}| = r - |A_i|$  by induction on  $k$ . For  $k = 1$  we have that  $f_{1,r}^{A_1} = 2$  and  $f_{1,r}^{A_2} = 1$  hold for all  $r$  (see Figure 7). Thus  $2 = f_{1,r}^{A_1} \leq f_{1,r}^{A_2} + f_{1,r}^{A_2} = 1 + 1$  and therefore (18) is true for  $k = 1$ , which completes the base

case of the induction. Assume now that for all  $i = 2, \dots, r$ , (18) holds for  $k$ . By (3) and (18), for all  $A_i$ ,  $f_{k+1,r}^{A_i}$  becomes

$$f_{k+1,r}^{A_i} = \min \begin{cases} f_{k,r}^{A_i \cup S} + f_{k,r}^{A_i \cup \widehat{S}} & S \cap \widehat{S} = \emptyset, S \cap A_i = \emptyset, \widehat{S} \cap A_i = \emptyset \\ & \text{and } |S| + |\widehat{S}| = r - |A_i|. \\ f_{k,r}^{A_{i-1}} & a \in A_{i-1}, A_{i-1} \subset A_i \text{ and } |A_{i-1}| = i - 1. \end{cases} \quad (19)$$

$$\begin{aligned} & \text{where the minimum is taken over all } A_{i-1}, S, \widehat{S} \\ & = f_{k,r}^{A_{i-1}}. \end{aligned} \quad (20)$$

Then for all  $2 \leq i \leq r$  and for all  $S, \widehat{S} \subseteq R : S \cap \widehat{S} = \emptyset, S \cap A_i = \emptyset, \widehat{S} \cap A_i = \emptyset$  and  $|S| + |\widehat{S}| = r - |A_i|$ , we can apply (20) to  $A_i \cup S$  and  $A_i \cup \widehat{S}$ , respectively, and get:

$$\begin{aligned} f_{k+1,r}^{A_i \cup S} + f_{k+1,r}^{A_i \cup \widehat{S}} &= f_{k,r}^{A_{i-1} \cup S} + f_{k,r}^{A_{i-1} \cup \widehat{S}} && \text{by (20)} \\ &\geq f_{k,r}^{A_{i-2}} && \text{by (18)} \\ &= f_{k+1,r}^{A_{i-1}} && \text{by (20)}. \end{aligned}$$

Note that the first equation is true because as for  $A_i \cap S = \emptyset$ , we have  $|A_{i-1} \cup S| = |A_i \cup S| - 1$ . The same holds for  $\widehat{S}$ . So altogether, we now have that (18) holds for  $k + 1$ .

Applying (18) to (19) gives  $f_{k+1,r}^{A_i} = f_{k,r}^{A_{i-1}}$  for all  $2 \leq i \leq r$ . So now we have that (20) holds for all  $k + 1 \geq i - 1$ , and thus we have  $f_{k,r}^{A_{i-1}} = f_{k-1,r}^{A_{i-2}} = f_{k-2,r}^{A_{i-3}} = \dots = f_{k-(i-2),r}^{A_1} = f_{k-(i-2),r}^{A_1} = f_{k-(i-2),r}$ , which completes the proof.  $\square$

*Proof (Theorem 4)* If  $\tilde{r} = r$ , we immediately have  $f_{k+1,r}^{A_i} = f_{k+1,\tilde{r}}^{A_i}$ , so there is nothing to show. So we now consider the case  $\tilde{r} < r$  and prove the statement by induction on  $k$ . For  $k = 0$ , we consider  $f_{1,r}^{A_i}$  and  $f_{1,\tilde{r}}^{A_i}$ . These values are defined only if  $|A_i| \leq 2^{0+1} = 2$  and therefore  $i = 1$  or  $i = 2$ . Thus we have to consider  $f_{1,r}^{A_1}, f_{1,\tilde{r}}^{A_1}, f_{1,r}^{A_2}$  and  $f_{1,\tilde{r}}^{A_2}$  for all  $2 \leq \tilde{r} \leq r$ . As can be seen in Figure 7, we have for all  $2 \leq \tilde{r} \leq r$ :  $f_{1,r}^{A_1} = 2 = f_{1,\tilde{r}}^{A_1}$  and  $f_{1,r}^{A_2} = 1 = f_{1,\tilde{r}}^{A_2}$ . This completes the base case of the induction for all  $i \in \{1, \dots, \tilde{r}\}$ .

Now assume Theorem 4 holds for  $k$ , and consider  $k + 1$ . We start with the case  $i \in \{2, \dots, \tilde{r}\}$  and find:

$$\begin{aligned} f_{k+2,\tilde{r}}^{A_i} &= f_{k+1,\tilde{r}}^{A_{i-1}} && \text{by Theorem 3} \\ &\geq f_{k+1,r}^{A_{i-1}} && \text{as Theorem 4 holds for } k \\ &= f_{k+2,r}^{A_i} && \text{by Theorem 3.} \end{aligned}$$

So Theorem 4 holds for  $i \in \{2, \dots, \tilde{r}\}$ . Now consider the case  $i = 1$ :

$$\begin{aligned}
f_{k+2,\tilde{r}}^{A_1} &= \min\{f_{k+1,\tilde{r}}^{A_1} + f_{k+1,\tilde{r}}^{A_{\tilde{r}}}, f_{k+1,\tilde{r}}^{A_2} + f_{k+1,\tilde{r}}^{A_{\tilde{r}-1}}, f_{k+1,\tilde{r}}^{A_3} + f_{k+1,\tilde{r}}^{A_{\tilde{r}-2}}, \dots, F_3\} \quad \text{by (4)} \\
&\quad \text{where } F_3 := \begin{cases} f_{k+1,\tilde{r}}^{A_{\tilde{p}}} + f_{k+1,\tilde{r}}^{A_{\tilde{p}+1}} & \text{if } \tilde{r} = 2\tilde{p} \\ 2 \cdot f_{k+1,\tilde{r}}^{A_{\tilde{p}}} & \text{if } \tilde{r} = 2\tilde{p} - 1 \end{cases} \\
&\geq \min\{f_{k+1,r}^{A_1} + f_{k+1,r}^{A_{\tilde{r}}}, f_{k+1,r}^{A_2} + f_{k+1,r}^{A_{\tilde{r}-1}}, f_{k+1,r}^{A_3} + f_{k+1,r}^{A_{\tilde{r}-2}}, \dots, F_4\} \quad \text{as Theorem 4 holds for } k \\
&\quad \text{where } F_4 := \begin{cases} f_{k+1,r}^{A_{\tilde{p}}} + f_{k+1,r}^{A_{\tilde{p}+1}} & \text{if } r = 2p \\ 2 \cdot f_{k+1,r}^{A_{\tilde{p}}} & \text{if } r = 2p - 1 \end{cases} \\
&\geq \min\{f_{k+1,r}^{A_1} + f_{k+1,r}^{A_r}, f_{k+1,r}^{A_2} + f_{k+1,r}^{A_{r-1}}, f_{k+1,r}^{A_3} + f_{k+1,r}^{A_{r-2}}, \dots, F_4\} \quad \text{by Theorem 1} \\
&\quad \text{where } F_4 := \begin{cases} f_{k+1,r}^{A_p} + f_{k+1,r}^{A_{p+1}} & \text{if } r = 2p \\ 2 \cdot f_{k+1,r}^{A_p} & \text{if } r = 2p - 1 \end{cases} \\
&= f_{k+2,r}^{A_1} \quad \text{by (4)}.
\end{aligned}$$

Hence also in the case where  $i = 1$ , Theorem 4 holds. This completes the proof.  $\square$

*Proof (Lemma 2)* Let  $r = 2p \geq 2$ , i.e.  $p \geq 1$ .

1. We start with the case  $k \leq p$ . We have  $f_{1,r} = 2$  for all  $r$  (see Figure 7). By Theorem 2 we know that  $f_{k,r}$  is monotonically increasing in  $k$ , and thus  $2 = f_{1,r} \leq f_{p,r}$ . We now use the standard decomposition for  $T_p$  in order to derive its two maximal pending rooted subtrees  $T_{p-1}$  with roots  $\rho_1$  and  $\rho_2$ . Using the same construction as in the proof of Observation 1, which is depicted in Figure 5, we can achieve  $X_{\rho_1} = A_p = \{a, c_1, \dots, c_{p-1}\}$  and  $X_{\rho_2} = A'_p = \{a, c_p, \dots, c_{2p-2}\}$  by assigning  $a$  to one leaf in each subtree  $T_{p-1}$ , respectively, where  $R = \{a, c_1, \dots, c_{2p-1}\}$ , and such that all other subtrees use one state each which is unique to this subtree. Thus, the root of  $T_p$  will have the MP root state estimate  $A_p \cap A'_p = \{a\}$ . (Note that no leaf is assigned character state  $c_{2p-1}$ , so we do not even require all states. We will need this fact later.) So we conclude  $f_{p,r} \leq 2$ , which together with  $f_{p,r} \geq 2$  as explained above completes  $f_{p,r} = 2$ . So together with Theorem 2 we achieve that  $f_{k,r} = 2$  for all  $k = 1, \dots, p$ .
2. Now consider the case  $k = p + 1$ . In this case we have with (4)

$$\begin{aligned}
f_{p+1,r} &= f_{p+1,2p} = \min\{f_{p,2p} + f_{p,2p}^{A_{2p}}, f_{p,2p}^{A_2} + f_{p,2p}^{A_{2p-1}}, \dots, f_{p,2p}^{A_p} + f_{p,2p}^{A_{p+1}}\} \\
&= \min\{f_{p,2p} + f_{p,2p}^{A_{2p}}, f_{p-1,2p} + f_{p,2p}^{A_{2p-1}}, \dots, f_{1,2p} + f_{0,2p}\} \quad \text{by Theorem 3} \\
&= \min\{2 + f_{p,2p}^{A_{2p}}, 2 + f_{p,2p}^{A_{2p-1}}, \dots, 2 + 1\} \quad \text{by Lemma 2, part 1} \\
&= 3.
\end{aligned}$$

The latter equation is true because  $1 \leq f_{p,2p}^{A_{2p}} \leq f_{p,2p}^{A_{2p-1}} \leq \dots \leq f_{p,2p}^{A_{p+1}}$  (at least one leaf has to be labelled  $a$  if  $a$  shall appear in the MP root state estimate).

3. Now consider the case  $k = r$ . We can proceed as above and assign  $X_{\rho_1} = A_p = \{a, c_1, \dots, c_{p-1}\}$  to the first of the two  $T_{r-1}$  subtrees induced by the standard decomposition, and  $X_{\rho_2} = A'_{p+1} = \{a, c_p, \dots, c_{2p-1}\}$  to the other one, where again  $R = \{a, c_1, \dots, c_{2p-1}\}$ . Then by (4) we can conclude  $f_{r,r} \leq f_{r-1,r}^{A_p} + f_{r-1,r}^{A_{p+1}}$ . Note that  $f_{r-1,r}^{A_p} + f_{r-1,r}^{A_{p+1}} = f_{p,r} + f_{p-1,r}$  by Theorem 3, so that  $f_{r,r} \leq f_{p,r} + f_{p-1,r} = 2 + 2$ , where the latter equation holds because of the first part of Lemma 2. So altogether,  $f_{r,r} \leq 4$ . By the monotonicity of Theorem 2, we obtain  $f_{k,r} \leq 4$  for all  $p + 1 < k \leq r$ .

$\square$

## References

1. M. Steel and M. Charleston, Five surprising properties of parsimoniously colored trees, *Bulletin of Mathematical Biology*, Volume 57, page 367-375 (1995)
2. W. M. Fitch, Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology, *Systematic Zoology*, Volume 20, page 406-416 (1971)
3. L. Herbst, Ancestral state reconstruction with parsimony, Master Thesis, Greifswald (2015)
4. I. P. Goulden and D. M. Jackson, *Combinatorial Enumeration*, Wiley, New York (1983)
5. C. Semple and M. Steel, *Phylogenetics*, Oxford University Press, New York (2003)
6. J. Felsenstein, *Inferring Phylogenies*, Sinauer Associates, Inc (2004)
7. D. A. Liberles (ed), *Ancestral Sequence Reconstruction*, Oxford University Press, New York (2007)
8. L. Szekely, P. Erdős and M. Steel, The Combinatorics of Evolutionary Trees - a Survey, *Journal of Combinatorial Mathematics and Combinatorial Computing*, Volume 15, page 241-254 (1994)
9. O. Gascuel and M. Steel, Predicting the ancestral character changes in a tree is typically easier than predicting the root state, *Systematic Biology*, Volume 63, page 421-435 (2014)
10. O. Gascuel and M. Steel, Inferring ancestral sequences in taxon-rich phylogenies, *Mathematical Biosciences*, Volume 227, page 125-153 (2010)
11. C. Semple and M. Steel, Tree reconstruction from multi-state characters, *Advances in Applied Mathematics*, Volume 28, page 169-184 (2002)
12. G. Li, M. Steel and L. Zhang, More taxa are not necessarily better for the reconstruction of ancestral character states, *Systematic Biology*, Volume 57, page 647-653 (2008)
13. W. Cai, J. Pei and N. V. Grishin, Reconstruction of ancestral protein sequences and its applications, *BMC Evolutionary Biology*, Volume 4, page 33 (2004)
14. O. W. Griffith, D. G. Blackburn, M. C. Brandley, J. U. Van Dyke, C. M. Whittington, M. B Thompson, Ancestral state reconstructions require biological evidence to test evolutionary hypotheses: A case study examining the evolution of reproductive mode in squamate reptiles, *Journal of Experimental Zoology*, Volume 324, page 493-503 (2015)
15. M. Fischer, V. Liebscher: On the balance of unrooted trees. 2015. Preprint. arXiv:1510.07882.

